Logistic Regression

Process

Decide on a **model**

Find the function which fits the data best

Choose a loss function

Pick the function which minimizes loss on data

Use function to make prediction on new examples

Logistic Regression

Actually classification, not regression :)

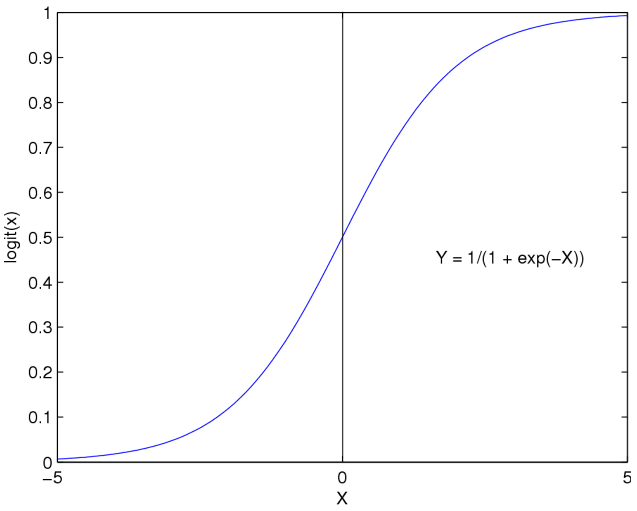
Learn $\mathbb{P}(Y=1|X=x)$ using $\sigma(w^T x)$, for link function $\sigma =$

Logistic function(or Sigmoid): $\frac{1}{1+exp(-z)}$

$$\mathbb{P}[Y=1|X=x,w] = \sigma(w^T x) = \frac{1}{1+\exp(-w^T x)}$$

$$\mathbb{P}[Y=0|X=x,w] = 1-\sigma(w^T x) = \frac{\exp(-w^T x)}{1+\exp(-w^T x)}$$

$$= \frac{1}{1+\exp(w^T x)}$$



Features can be discrete or continuous!

Sigmoid for binary classes

$$\mathbb{P}(Y = 0|w, X) = \frac{1}{1 + \exp(w_0 + \sum_k w_k X_k)}$$

$$\mathbb{P}(Y = 1|w, X) = 1 - \mathbb{P}(Y = 0|w, X) = \frac{\exp(w_0 + \sum_k w_k X_k)}{1 + \exp(w_0 + \sum_k w_k X_k)}$$

$$\frac{\mathbb{P}(Y = 1|w, X)}{\mathbb{P}(Y = 0|w, X)} =$$

Sigmoid for binary classes

$$\mathbb{P}(Y=0|w,X) = \frac{1}{1+\exp(w_0 + \sum_k w_k X_k)}$$

$$\mathbb{P}(Y=1|w,X) = 1 - \mathbb{P}(Y=0|w,X) = \frac{\exp(w_0 + \sum_k w_k X_k)}{1+\exp(w_0 + \sum_k w_k X_k)}$$

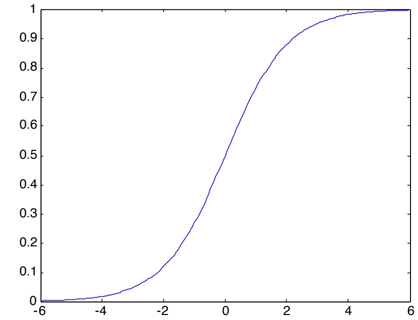
$$\frac{\mathbb{P}(Y=1|w,X)}{\mathbb{P}(Y=0|w,X)} = \exp(w_0 + \sum_k w_k X_k)$$

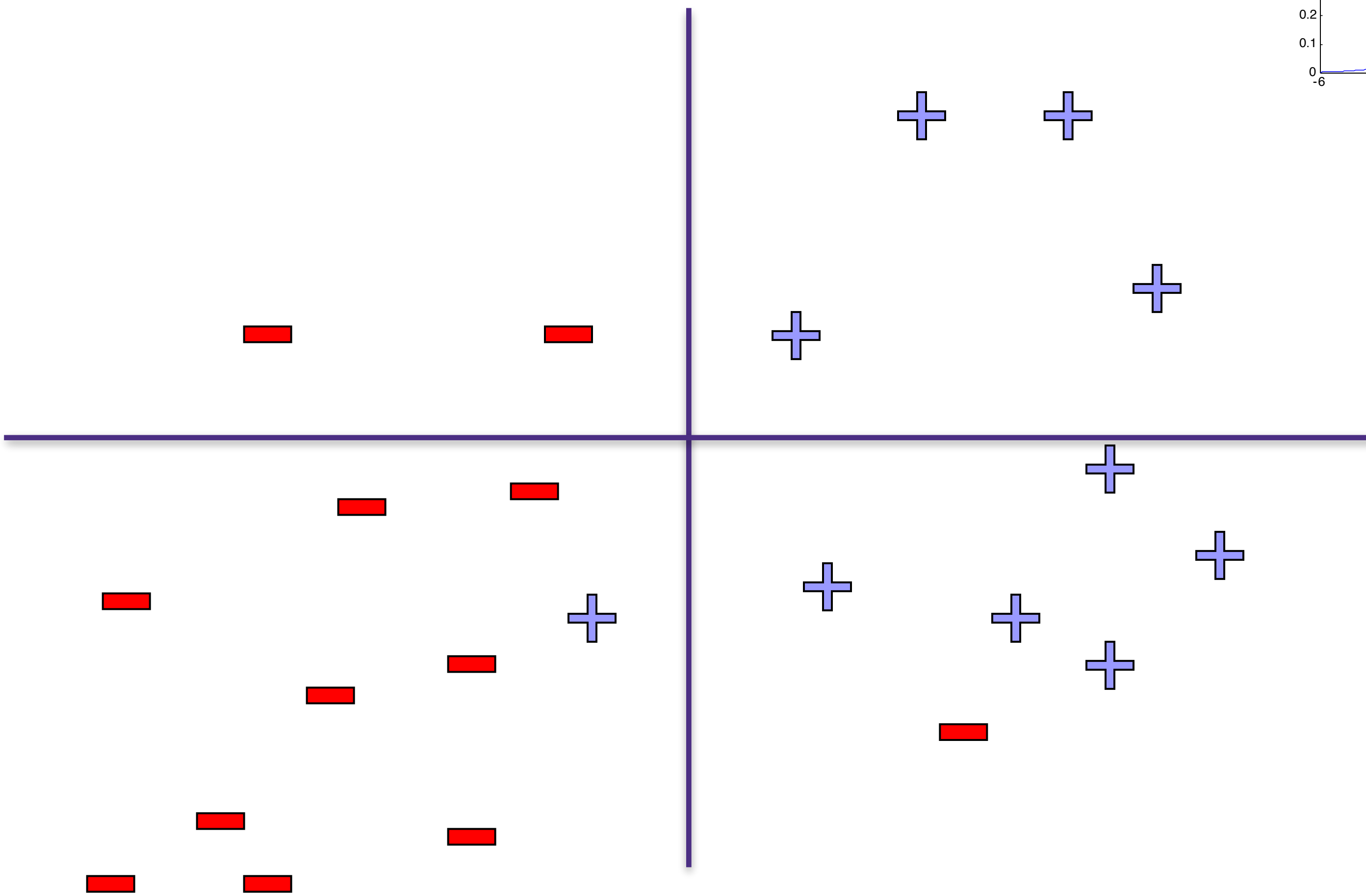
Linear Decision Rule!

$$\log \frac{\mathbb{P}(Y=1|w,X)}{\mathbb{P}(Y=0|w,X)} = w_0 + \sum_k w_k X_k$$

Logistic Regression – a Linear classifier

$$\frac{1}{1+exp(-z)}$$





$$\ln \frac{P(Y=0|X)}{P(Y=1|X)} = w_0 + \sum_i w_i X_i$$

Process

Decide on a **model**

Find the function which fits the data best

- **Choose a loss function**
- **Pick the function which minimizes loss on data**

Use function to make prediction on new examples

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y=-1|x,w) = \frac{1}{1+\exp(w^T x)}$$

$$P(Y=1|x,w) = \frac{\exp(w^T x)}{1+\exp(w^T x)}$$

- **This is equivalent to:**

$$P(Y=y|x,w) = \frac{1}{1+\exp(-y\, w^T x)}$$

- **So we can compute the maximum likelihood estimator:**

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$

$$= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w))$$

Logistic Loss: $\ell_i(w) = \log(1 + \exp(-y_i\, x_i^T w))$

Squared error Loss: $\ell_i(w) = (y_i - x_i^T w)^2$

(MLE for Gaussian noise)

Process

Decide on a **model**

Find the function which fits the data best

Choose a loss function

Pick the function which minimizes loss on data

Use function to make prediction on new examples

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$

$$= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w)) = J(w)$$

What does $J(w)$ look like? Is it convex?

Loss function: Conditional Likelihood

Loss function: Conditional Likelihood

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$

$$= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w)) = J(w)$$

Good news: $J(\mathbf{w})$ is convex function of **w**, no local optima problems

Bad news: no closed-form solution to maximize $J(\mathbf{w})$

Good news: convex functions easy to optimize

One other concern… overfitting.

- **Have a bunch of iid data:** $\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

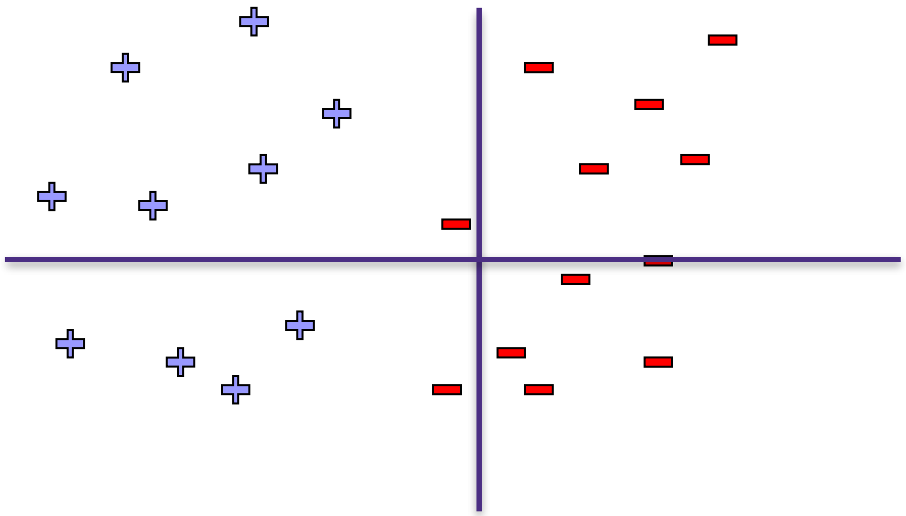
$$\widehat{w}_{MLE} = \arg\max_w \prod_{i=1}^n P(y_i|x_i, w)$$
$$= \arg\min_w \sum_{i=1}^n \log(1 + \exp(-y_i\, x_i^T w))$$

Does anyone see a situation when this minimization might overfit?

Overfitting and Linear Separability

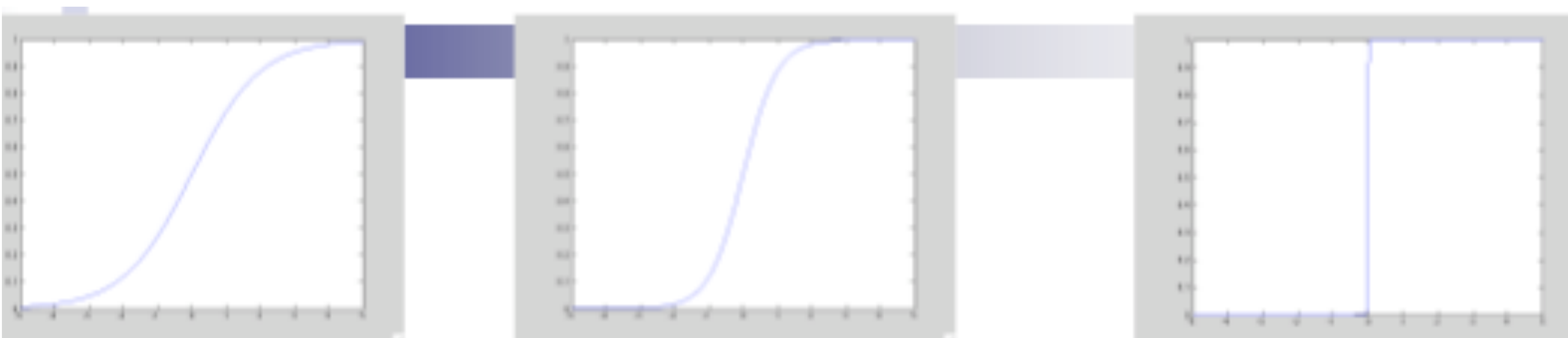
$$\arg\min_{w} \sum_{i=1}^{n} \log(1 + \exp(-y_i \, x_i^T w))$$

When is this loss small?



Large parameters → Overfitting

When data is linearly separable, weights $\Rightarrow \infty$



$$\frac{1}{1+e^{-x}} \qquad \frac{1}{1+e^{-2x}} \qquad \frac{1}{1+e^{-100x}}$$

Overfitting

Penalize high weights to prevent overfitting?

Regularized Conditional Log Likelihood

Add a penalty to avoid high weights/overfitting?:

$$\arg\min_{w,b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i\left(x_i^T w + b\right))\right) + \lambda ||w||_2^2$$

Be sure to not regularize the offset b!